Információkinyerés magyar nyelvű önéletrajzokból a nexum Karrierportálhoz

Farkas Richárd[1], Dobó András[3], Kurai Zoltán[3], Miklós István[1], Miszori Attila[3], Nagy Ágoston[1], Vincze Veronika[2], Zsibrita János[3]

[1] Szegedi Tudományegyetem, Informatikai Tanszékcsoport
[2] MTA-SZTE Mesterséges Intelligencia Kutatócsoport
[3] nexum Magyarország Kft.

Kapcsolat: rfarkas@inf.u-szeged.hu

Kivonat

Számos nagyvállalatnak komoly problémát jelent az alkalmazottak toborzása. A legnagyobb gondot az jelenti, hogy akkora mennyiségben jelentkeznek ezekhez a cégekhez egy adott kiírásra, hogy nincs arra elegendő emberi erőforrás, hogy a rengeteg, sokszor több ezernyi beérkezett önéletrajzot egyesével végignézzék. Ezért a bevett szokás az, hogy az adatbázisban lévő önéletrajzok közül tulajdonképpen véletlenszerűen választanak pozíciótól függően néhány tízet vagy néhány százat, mert csak ezek végigolvasására van idejük.

A Szegedi Tudományegyetem Nyelvtechnológiai Csoportja és a nexum Magyarország kft. közös kutatás-fejlesztési projektje során egy olyan módszer kifejlesztésén dolgozunk, mely egy adott álláslehetőséghez megadott lekérdezést és egy önéletrajzhalmazt inputként kapva visszaadja az önéletrajzok rendezett sorozatát az illetők adott pozícióra való alkalmassága alapján. Mivel a gyakorlatban az látszik, hogy minden önéletrajz egyedileg szerkesztett, mindegyik más és más struktúrájú, ezért az önéletrajzok megfelelőségük szerint közvetlenül nem rangsorolhatók. Ahhoz, hogy ez megvalósítható lehessen, szükség van arra, hogy a munkavállalók adatait egy egységes adatstruktúrába ki tudjuk nyerni az önéletrajzokból.

Ezért első lépésként egy olyan módszert fejlesztettünk ki, mely alkalmas arra, hogy egy tetszőleges önéletrajzból a munkavállaló legfontosabb adatait, mint például a nevét, a születési dátumát, az elérhetőségeit, a tanulmányi adatait, a munkatapasztalatait, a nyelvismeretét és további lényeges adatait kinyerje. A legtöbb nagyvállalat karrierportáljában az önéletrajz beadása mellett a munkakeresőknek egy űrlapot is ki kell tölteniük, melyen az önéletrajzi adataikat strukturáltan megadják. Mivel az algoritmusunk alkalmas arra, hogy ezeket az adatokat az önéletrajzból kinyerje, ezért az űrlapkitöltési folyamat automatizálásában is fel lehetne használni módszerünket úgy, hogy a munkakeresőnek az adatait csak ellenőriznie kelljen, és ne neki kelljen mindent manuálisan bevinnie az űrlapba.

Ennek a feladatnak az első problémáját az okozta, hogy a munkavállalók önéletrajzaikat rendszerint különböző fájlformátumban küldik be a nagyvállalatok karrierportálján keresztül. Hogy algoritmusunk egységes formátumú önéletrajzokat kapjon bemenetként, a különböző formátumokból először egységesen PDF formátumot ké-

szítettünk, majd a PDF formátumú önéletrajzokból egyszerű szöveges dokumentumokat gyártottunk. Ezek a szöveges dokumentumok a formázásokat ugyan mellőzik, de a dokumentumok elrendezését megtartják.

A feladat ezek után a célinformáció kinyerése a szöveges (de strukturált) állományokból. Kézzel annotált tanító dokumentumok hiányában azt a megoldást láttuk kézenfekvőnek, hogy megpróbálunk egy olyan módszert kidolgozni, mellyel tanító adatok automatikusan generálhatók. Ez az álláskeresők által a karrierportálon az önéletrajz feltöltése mellett kitöltött űrlapok alapján megvalósítható. Mivel az adatok az űrlapon strukturáltan kerültek felvitelre, és elméletileg ugyanazok az adatok szerepelnek az önéletrajzban is, ezért az űrlap adatainak önéletrajzra való mappelésével automatikus tanító önéletrajzokat kaphatunk.

Rendszerünk az önéletrajzok előfeldolgozása – amely magában foglalja a szöveg normalizálását és a dokumentumok struktúrájának egy belső fareprezentációba történő illesztését – után automatikusan tanító adatokat generál az önéletrajz-űrlap párosokból. Habár kezdetben ez viszonylag egyszerű feladatnak tűnt, számos problémával kellett szembenéznünk. Először is, rengeteg önéletrajz-formátum, -struktúra fordul elő a beadott önéletrajzok között, és sok egyáltalán nincs is vagy csak alig van strukturálva. Másodszor, bár úgy gondoltuk, hogy a feltöltött önéletrajz és a vele egy időben kitöltött űrlap adatai megegyeznek, valójában sok helyen különböznek, egyes adatok a két helyen különböző formában szerepelnek, illetve sok adat pusztán az egyik helyen szerepel. Ezen kívül a rengeteg elgépelés is nagyban nehezíti a munkát. Megoldásként az adatokat próbáltuk normalizálni, a felismerésben különféle mintákat használtunk, és a különböző adatosztályokhoz külön annotátorfüggvényeket készítettünk. Az így automatikusan generált tanító adatok mellett kézzel annotált dokumentumokat is felhasználtunk a tanításban.

A tanító adatok elkészítése után egy MEMM szekvenciajelölő modellt tanítunk [1], melyhez számos különféle jellemzőt definiáltunk, többek közt különféle reguláris kifejezéseket, listákat, szóalaki jellemzőket, mondat- és szövegbeli elhelyezkedést és a dokumentum struktúrájában elfoglalt pozíciót, kézzel gyűjtött doméntaxonómiákat stb. Az így tanított modell egy még annotálatlan önéletrajzot megkapva képes az önéletrajzban található fontosabb adatok jó minőségű kinyerésére. Természetesen a kinyerés minősége nagyban függ a kapott önéletrajz strukturáltságától és minőségétől is.

Demónkban lehetőség nyílik a rendszer megismerésére éles működés közben, továbbá a fejlesztés során megoldott gyakorlati nyelvtechnológiai problémák megvitatására.

Hivatkozások

1. McCallum, Andrew, Freitag, Dayne, Pereira, Fernando: Maximum Entropy Markov Models for Information Extraction and Segmentation. Proc. ICML (2000) 591–598